bioRxiv preprint doi: https://doi.org/10.1101/2020.10.12.335521; this version posted October 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Codon pattern reveals SARS-CoV-2 to be a monomorphic strain that emerged through recombination of replicase and envelope alleles of bat and pangolin origin

Kanika Bansal[1, *], Prabhu B. Patil[1, *]

[1]CSIR- Institute of Microbial Technology, Chandigarh, India

[*]Corresponding author

Address for correspondence:

Kanika Bansal, PhD
Research Associate
Laboratory of Bacterial Genomics and Evolution
CSIR- Institute of Microbial Technology, Chandigarh, India, 160036
Email: kanikabansal@imtech.res.in

Prabhu B. Patil, PhD
Principal Scientist
Laboratory of Bacterial Genomics and Evolution
CSIR- Institute of Microbial Technology, Chandigarh, India, 160036
Email: pbpatil@imtech.res.in

Abstract

Viruses are dependent on the host tRNA pool, and an optimum codon usage pattern (CUP) is a driving force in its evolution. Systematic analysis of CUP of replicase (rdrp), spike, envelope (E), membrane glycoprotein (M), and nucleocapsid (N) encoding genes of SARS-CoV-2 from reported diverse lineages to suggest one-time host jump of a SARS-CoV-2 isolate into the human host. In contrast to human isolates, a high degree of variation in CUP of these genes suggests that bats, pangolins, and dogs are natural reservoirs of diverse strains. At the same time, our analysis suggests that dogs are not a source of SARS-CoV-2. Interestingly, CUP of rdrp displays conservation with two bat SARS isolates RaTG13 and RmYN02. CUP of the SARS-CoV-2 E gene is also conserved with bat and pangolin isolates with variations for a few amino acids. This suggests role allele replacement in these two genes involving SARS strains of least two hosts. At the same time, a relatively conserved CUP pattern in replicase and envelope across hosts suggests them it to be an ideal target in antiviral development for SARS-CoV-2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.12.335521; this version posted October 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

The origin and success of novel SARS coronavirus (SARS-CoV-2) (betacoronavirus) causing the COVID-19 disease pandemic has been a topic of intense discussion. In the past two decades since the first outbreak of SARS in 2002, several SARS-related coronaviruses were reported from the bat, which was speculated to be significant reservoirs for the future possible outbreaks (Cui et al., 2019; Ge et al., 2013; Hu et al., 2017; Li et al., 2005). Bats are the only flying mammals representing 20% of known mammalian species and are critical natural reservoirs of many zoonotic viruses like Nipah virus, Hendra virus, rabies virus, Ebola virus, etc. (Halpin et al., 2000; Leroy et al., 2005; Mackenzie et al., 2001). Besides bat, a considerable number of wild animals have played a pivotal role in zoonotic transfers (Bengis et al., 2004). According to reports before the SARS-CoV-2 pandemic, there was a high-risk assessment of SARS coronavirus infection from wild animals like bats, civets, pangolins, snakes, tiger, and primates in China due to human interventions (Bell et al., 2004; Gottlieb, 2003; TANG et al., 2006).

The human-wildlife interface as a part of culture or globalization poses risks for zoonotic transfers followed by disease outbreaks like coronavirus outbreaks: SARS (2002,2003), MERS (2012), and SARS-CoV-2 (2019). Their genome similarities with already reported SARS viruses from diverse animals are estimated by animal reservoirs for such outbreaks. For instance, SARS 2003 outbreak virus had 99.6% genome similarity with palm civets indicating it to be a direct source. Just 0.4% divergence from the animal reservoir stipulates its recent transfer into the masked palm civet population (Shi and Hu, 2008). Despite genetic diversity with bat SARS-CoV they were ultimately found to be a source of the pandemic due to no pathogen prevalence in wild civet population and clinical symptom manifestation in civets, unlike bats (Li et al., 2005). However, in the current pandemic, there are several theories of the origin of SARS-CoV-2 either from bat, pangolin, dog or some intermediate host, etc. (Paraskevis et al., 2020; Zhang et al., 2020). The closest match to SARS-CoV-2 is RaTG13 (96% identity) isolated from the *Rhinolophus affinis* bat (Zhou et al., 2020), followed by pangolin SARS viruses with 91% identity (Zhou et al., 2020). As the closest match is just 96%, it has opened a heated debate in the scientific community for its origin, and no direct animal source can be detected.

According to genome similarities, SARS-CoV-2 differs from its closest SARS coronavirus by 4%, followed by 9% with its next closest relative, pangolin. It indicates that the virus has

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.12.335521; this version posted October 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

evolved before infecting humans, and there is a missing link between bat/pangolin and humans, which further inflates the argument on the animal source. Nevertheless, another study based on CpG island deficiency in SARS-CoV-2 and canine coronavirus (alphacoronavirus) suggested that dogs may have provided a cellular environment for SARS-CoV-2 evolution into a CpG deficient virus (Xia, 2020). Hence, they claim dog to be a direct source of the current pandemic, raising a constant debate (Pollock et al., 2020) (https://www.linkedin.com/pulse/where-dog-laymans-version-my-mbe-paper-xuhua-xia/). Nevertheless, most of the other RNA viruses like pestvirus, in addition to bat or pangolin SARS-CoV, are also depleted in CpG is not included in the study. Further, the spike receptor-binding domain of SARS-CoV-2 was similar to pangolin SARS strains compared to bat SARS. CpG island deficiency is not a unique feature of dog SARS-CoV, and Pollock et al. have concluded that there is no direct evidence for the role of dogs as intermediate hosts. Hence, we need to address this issue with other fundamental evidence.

For host jump events, viral codon optimization based on the host tRNA pool is critical (Khandia et al., 2019; Tian et al., 2018; Van Weringh et al., 2011). In the present study, we have focused on codon usage pattern (CUP) of SARS coronavirus from different hosts under debate (bat, pangolin, and dog) as a probable origin for SARS-CoV-2. Usage patterns of synonymous codons are a critical feature in the adaptation of organisms as viruses are dependent on the host tRNA pool for replication and disease manifestations. For instance, codon adaptation indices were studied for retroviruses infecting humans, including the HIV-1 virus (RoyChoudhury and Mukherjee, 2013). Once the viral genome is in the host translational mechanism, genes having optimized codons according to the host translate faster, resulting in higher fitness of the virus (Carbone, 2008). Hence, an optimum CUP is vital in its evolution, and probable host jumps. This also results in synonymous changes in the viral genome, which are not revealed by protein mutational studies. Codon usage study of SARS-CoV-2 has found high AU content influencing its codon usage and better adaptation to the humans (Dilucca et al., 2020). However, another study comparing the codon usage pattern of SARS-CoV-2 with other betacoronaviruses suggested that current pandemic coronavirus is subjected to different evolutionary pressures (Gu et al., 2020). However, systematic insights into CUP is required to understand its origin and phenomenal success of emergent viruses like SARS-CoV-2.

The genome of SARS-CoV-2 is a single positive-stranded RNA of approximately 30,000 nucleotides. Major structural proteins are spike protein (S), envelope protein (E), membrane

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.12.335521; this version posted October 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

glycoprotein (M), nucleocapsid protein (N), and non-structural RNA dependent RNA polymerase (rdrp). As these five viral proteins are common amongst betacoronaviruses (Woo et al., 2010), it will be useful in studying diverse coronaviruses from diverse hosts, i.e., humans, bat, pangolin, and dog. Presently, we have analyzed CUP of these five proteins in SARS coronavirus, which are considered important targets for vaccine and antiviral development for SARS-CoV-2 (Aftab et al., 2020; Du et al., 2009; Huang et al., 2020; Wu et al., 2020). CUP calculations should be based on highly redundant amino acids rather than amino acids having one or two codons. Here, we have calculated the percentage of GC biased synonymous codons for amino acids having at least four synonymous codons (Glycine, Valine, Threonine, Leucine, Arginine, Serine, Proline and Alanine) (Patil and Sonti, 2004). Patil et al. have proven how CUP can detect horizontally acquired genes from a diverse background under selection pressure by analyzing the codon usage pattern of each amino acid in a particular gene in a graphical way. Similarly, a host jump event may lead to codon optimization, reflected in the CUP. Ideally, for an organism, its crucial genes should have a similar pattern of CUP. Nevertheless, genes pivotal for viral host jump and disease manifestations like spike or rdrp may show deviation from the pattern. Hence, CUP graphs enable us to visually inspect the patterns of synonymous changes across diverse hosts and further reflecting selection pressure in each of the genes.

Results and Discussion

We have compared CUP for all five genes amongst 134 SARS-CoV-2 genomes representing all the sub-lineages from the two major lineages A and B (supplementary table 1) (Rambaut et al., 2020). Here, genes having anonymous nucleotides were omitted from the analysis. CUPs of all the genes were uniform for all SARS-CoV-2 reported from the human population worldwide irrespective of ancestral or most recent lineages (supplementary figures 1, 2, 3, 4, and 5), indicating a single event of a zoonotic transfer of a thriving strain of SARS. However, each of them all the genes have distinct CUP among themselves, indicating distinct selection pressure (figure 1).

Since horseshoe bat SARS strains RaTG13 and RmYN02 are known to be the closest strains to SARS-CoV-2, we have compared CUP of these two strains with SARS-CoV-2 (figure 2). Here, the rdrp and E patterns overlapped for all the three strains with slight variations in E for serine and valine. However, all three strains have distinct patterns for S, M, and N genes. This reaffirms the previous studies that RaTG13 and RmYN02 are close to SARS-CoV-2

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.12.335521; this version posted October 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Zhou et al., 2020), indicating that SARS-CoV-2 is distinct from them. Unique S, M, and N patterns indicate SARS-CoV-2 has evolved to adapt to humans and to look for the possibility of intermediate hosts. We analyzed other known SARS viruses from hypothesized hosts.

Unlike SARS-CoV-2 isolates of human, CUP is variable in isolates of non-human hosts like a bat, pangolin, and dog (supplementary table 2) depicting ongoing adaptation and evolution of SARS in these hosts (figure 3). Interestingly, variations in CUP of S, M, and N genes are more marked across non-human hosts than the rdrp and E gene. For instance, rdrp CUP for pangolin and dog does not have that degree of variation and is host-specific compared to bat rdrp. Bat CUP for rdrp is variable in all the strains under study, correlating with the fact that bat is a reservoir of SARS coronaviruses. Whereas, CUP of E is conserved in bat and pangolin.

Interestingly, rdrp CUP of SARS-CoV-2 shows a slight overlap in pattern with all pangolin SARS strains, but it is not as prominent as two bat strains (figure 2). Among bat-SARS, pattern matches with just two previously suggested SARS, i.e., RaTG13 and RmYN02. In contrast, SARS-CoV-2 CUP of E has an overlapping pattern with all isolates of bat except for JTMC15 and serine and valine CUP. Interestingly, the precisely same pattern in all pangolin strains overlapping with SARS-CoV-2 except for valine is also observed. Unlike rdrp and E CUP of SARS-CoV-2, S, M, and N CUP does not have any similarity with non-human hosts studied till now. Further, no similarities in CUP of all five genes with dogs overrules the role of dog in SARS-CoV-2 evolution.

The above analysis reveals single patterns for all five genes in different lineages of SARS-CoV-2 affirms a single event of host jump of codon-optimized SARS strain from its animal reservoir. However, amongst five SARS-CoV-2 distinct CUP patterns are observed, pointing towards different evolutionary forces acting in the structural and non-structural genes. While comparing CUP of the current pandemic with all the hypothesized hosts, rdrp and E gene sources were found to be bat RaTG13 and RmYN02 and not for S, M, and N genes. This contradicts the previous study correlating rdrp, S, and N proteins with RaTG13 isolate (Gu et al., 2020). Though RaTG13 is the closest match to the current pandemic, some genes of SARS-CoV-2 does not share CUP with it. This indicates codon optimization in SARS-CoV-2, according to the new host (human) as compared to its animal reservoir.

Further, no gene has any overlap with dog coronavirus strains, ruling out the possibility of dog as an animal reservoir. Nevertheless, the unknown S, M, and N genes play pivotal roles

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.12.335521; this version posted October 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in host interactions and disease manifestations. Gene adaptations in S, M, and N have been pivotal in the emergence of the SARS viral pool in the hosts and the current pandemic. Suggesting that S, M, and N have mainly been under drastic variation both within and across the hosts. Either hyper-recombination events or selection pressure can explain this disparity in CUP conservation for rdrp, E than S, M, N genes. Hence, for the high variability across hosts, the mutation is not a rampant force (Boni et al., 2020). We can conclude that evolution in host interaction steps (S, M, and N) has more significance than at the level of replicability (rdrp) and packaging (E). In addition to this, out of five genes in bat, only E shows a unified CUP, overlapping with pangolin and SARS-CoV-2. This indicates that rdrp and E are under purifying selection and could be better targeted for antivirals development. High selection pressure in S, M, and N genes indicate that antiviral development attempts targeting these may be futile.

For centuries, bats are exploited as medicines, food products in southern China and Asia (Mickleburgh et al., 2002). Hence, wildlife-human interactions have been critical for the evolution of human SARS coronavirus outbreaks. The high diversity of CUP in bats compared to humans and pangolins is reflected by the high genetic diversity of SARS in bats (Li et al., 2005). Further, some of the SARS-related coronaviruses are found to have broad species tropism (Ge et al., 2013). Implicates that bats have provided a playground for SARS to mutate or recombine, resulting in a strain that might be successful in other hosts (in this case, humans) than bats. As the resulting strain (SARS-CoV-2) might be an accidental strain that originated from recombination in bat but was unfit in bat or its unknown animal reservoir, not leaving its evolutionary footprint. A highly similar strain to the present pandemic could not be found and further warrants studies to explore the diversity of SARS-like coronaviruses.

Methods

For each gene the frequency of codon usage for different amino acids was calculated using a web based program (www.bioinformatics.org). Further, eight amino acids i.e., Glycine, Valine, Threonine, Leucine, Arginine, Serine, Proline and Alanine that have atleast four synonymous codons were selected and the percentage of synonymous codons that end with G or C was calculated for each amino acid and gene. The pattern was calculated for a group of genes by plotting mean values ± SD corresponding to a particular amino acid.

References

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.12.335521; this version posted October 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Aftab, S.O., Ghouri, M.Z., Masood, M.U., Haider, Z., Khan, Z., Ahmad, A., and Munawar, N. (2020). Analysis of SARS-CoV-2 RNA-dependent RNA polymerase as a potential therapeutic drug target using a computational approach. Journal of translational medicine *18*, 1-15.

Bell, D., Roberton, S., and Hunter, P.R. (2004). Animal origins of SARS coronavirus: possible links with the international trade in small carnivores. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences *359*, 1107-1114.

Bengis, R., Leighton, F., Fischer, J., Artois, M., Morner, T., and Tate, C. (2004). The role of wildlife in emerging and re-emerging zoonoses. Revue scientifique et technique-office international des epizooties *23*, 497-512.

Boni, M.F., Lemey, P., Jiang, X., Lam, T.T.-Y., Perry, B., Castoe, T., Rambaut, A., and Robertson, D.L. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. bioRxiv.

Carbone, A. (2008). Codon bias is a major factor explaining phage evolution in translationally biased hosts. Journal of Molecular Evolution *66*, 210-223.

Cui, J., Li, F., and Shi, Z.-L. (2019). Origin and evolution of pathogenic coronaviruses. Nature Reviews Microbiology *17*, 181-192.

Dilucca, M., Forcelloni, S., Georgakilas, A.G., Giansanti, A., and Pavlopoulou, A.J.V. (2020). Codon Usage and Phenotypic Divergences of SARS-CoV-2 Genes. *12*, 498.

Du, L., He, Y., Zhou, Y., Liu, S., Zheng, B.-J., and Jiang, S. (2009). The spike protein of SARS-CoV—a target for vaccine and therapeutic development. Nature Reviews Microbiology *7*, 226-236.

Ge, X.-Y., Li, J.-L., Yang, X.-L., Chmura, A.A., Zhu, G., Epstein, J.H., Mazet, J.K., Hu, B., Zhang, W., and Peng, C. (2013). Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature *503*, 535-538.

Gottlieb, S. (2003). Chinese scientists must test wild animals to find the host of SARS (British Medical Journal Publishing Group).

Gu, H., Chu, D.K., Peiris, M., and Poon, L.L.J.V.E. (2020). Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. *6*, veaa032.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.12.335521; this version posted October 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Halpin, K., Young, P.L., Field, H., and Mackenzie, J. (2000). Isolation of Hendra virus from pteropid bats: a natural reservoir of Hendra virus. Journal of General Virology *81*, 1927-1932.

Hu, B., Zeng, L.-P., Yang, X.-L., Ge, X.-Y., Zhang, W., Li, B., Xie, J.-Z., Shen, X.-R., Zhang, Y.-Z., and Wang, N. (2017). Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS pathogens *13*, e1006698.

Huang, J., Song, W., Huang, H., and Sun, Q. (2020). Pharmacological therapeutics targeting RNA-dependent RNA polymerase, proteinase and spike protein: from mechanistic studies to clinical trials for COVID-19. Journal of Clinical Medicine *9*, 1131.

Khandia, R., Singhal, S., Kumar, U., Ansari, A., Tiwari, R., Dhama, K., Das, J., Munjal, A., and Singh, R.K. (2019). Analysis of Nipah virus codon usage and adaptation to hosts. Frontiers in microbiology *10*, 886.

Leroy, E.M., Kumulungui, B., Pourrut, X., Rouquet, P., Hassanin, A., Yaba, P., Délicat, A., Paweska, J.T., Gonzalez, J.-P., and Swanepoel, R. (2005). Fruit bats as reservoirs of Ebola virus. Nature *438*, 575-576.

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J.H., Wang, H., Crameri, G., Hu, Z., and Zhang, H. (2005). Bats are natural reservoirs of SARS-like coronaviruses. Science *310*, 676-679.

Mackenzie, J., Chua, K., Daniels, P., Eaton, B., Field, H., Hall, R., Halpin, K., Johansen, C., Kirkland, P., and Lam, S. (2001). Emerging viral diseases of Southeast Asia and the Western Pacific. Emerging infectious diseases *7*, 497.

Mickleburgh, S.P., Hutson, A.M., and Racey, P.A. (2002). A review of the global conservation status of bats. Oryx *36*, 18-34.

Paraskevis, D., Kostaki, E.G., Magiorkinis, G., Panayiotakopoulos, G., Sourvinos, G., Tsiodras, S.J.I., Genetics, and Evolution (2020). Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *79*, 104212.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.12.335521; this version posted October 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Patil, P.B., and Sonti, R.V. (2004). Variation suggestive of horizontal gene transfer at a lipopolysaccharide (lps) biosynthetic locus in Xanthomonas oryzae pv. oryzae, the bacterial leaf blight pathogen of rice. BMC microbiology *4*, 40.

Pollock, D.D., Castoe, T.A., Perry, B.W., Lytras, S., Wade, K.J., Robertson, D.L., Holmes, E.C., Boni, M.F., Kosakovsky Pond, S.L., and Parry, R. (2020). Viral CpG deficiency provides no evidence that dogs were intermediate hosts for SARS-CoV-2. Molecular biology and evolution.

Rambaut, A., Holmes, E., O'Toole, Á., Hill, V., McCrone, J., and Ruis, C.J.P.J. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020. *15*.

RoyChoudhury, S., and Mukherjee, D. (2013). Complex codon usage pattern and compositional features of retroviruses. Computational and mathematical methods in medicine *2013*.

Shi, Z., and Hu, Z. (2008). A review of studies on animal reservoirs of the SARS coronavirus. Virus research *133*, 74-87.

TANG, H.-w., HUANG, J.-c., and HE, J.-f. (2006). Analysis of risk factors of SARS coronaryvirus infection among population contacted with wild animals in Guangdong Province. China Tropical Medicine, 3.

Tian, L., Shen, X., Murphy, R.W., and Shen, Y. (2018). The adaptation of codon usage of+ ssRNA viruses to their hosts. Infection, Genetics and Evolution *63*, 175-179.

Van Weringh, A., Ragonnet-Cronin, M., Pranckeviciene, E., Pavon-Eternod, M., Kleiman, L., and Xia, X. (2011). HIV-1 modulates the tRNA pool to improve translation efficiency. Molecular biology and evolution *28*, 1827-1834.

Woo, P.C., Huang, Y., Lau, S.K., and Yuen, K.-Y.J.v. (2010). Coronavirus genomics and bioinformatics analysis. *2*, 1804-1820.

Wu, D., Koganti, R., Lambe, U.P., Yadavalli, T., Nandi, S.S., and Shukla, D.J.J.o.c.m. (2020). Vaccines and therapies in development for SARS-CoV-2 infections. *9*, 1885.

Xia, X. (2020). Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. Molecular biology and evolution.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.12.335521; this version posted October 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Zhang, T., Wu, Q., and Zhang, Z.J.C.B. (2020). Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., and Huang, C.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. nature *579*, 270-273.

Figure legends

Figure 1: Codon usage pattern for rdrp, S, E, M, and N genes of SARS-CoV-2 (reference strain, NC_045512.2). Eight amino acids with at least four synonymous codons are represented on the X-axis. The percentage of codons ending with G/C for each amino acid is represented on Y-axis.

Figure 2: CUP of rdrp, S, E, M, and N genes for human SARS-CoV-2 reference strain (NC_045512.2) is denoted by the dotted black line, and bat SARS isolates RaRG13 and RmYN02. Eight amino acids with at least four synonymous codons are represented on the X-axis. The percentage of codons ending with G/C for each amino acid is represented on Y-axis.

Figure 3: CUP for rdrp, S, E, M, and N genes for bat, pangolin, and dog. Here, CUP for reference strain of SARS-CoV-2 is represented by dotted lines. Eight amino acids with at least four synonymous codons are represented on the X-axis. The percentage of codons ending with G/C for each amino acid is represented on Y-axis.

Supplementary material

Supplementary table 1: Metadata for the human SARS-CoV-2 genomes used in the study

Supplementary table 2: SARS strains used in the present study.

Supplementary figure 1: CUP for the rdrp gene for the human SARS-CoV-2 isolates. Eight amino acids with at least four synonymous codons are represented on the X-axis. The percentage of codons ending with G/C for each amino acid is represented on Y-axis.

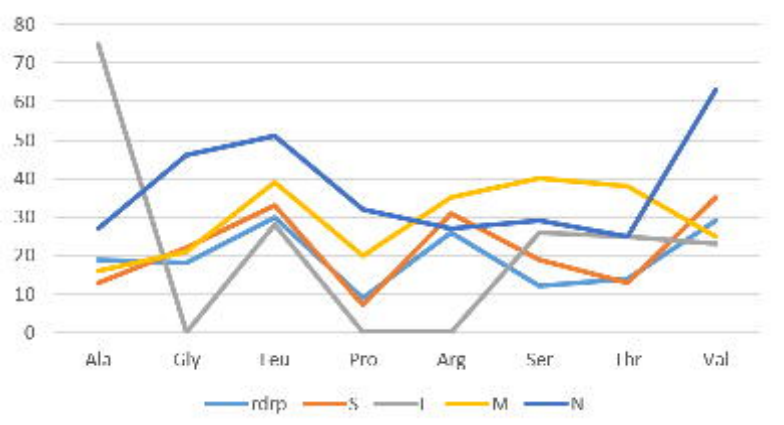
Supplementary figure 2: CUP for the spike gene for the human SARS-CoV-2 isolates. Eight amino acids with at least four synonymous codons are represented on the X-axis. The percentage of codons ending with G/C for each amino acid is represented on Y-axis.

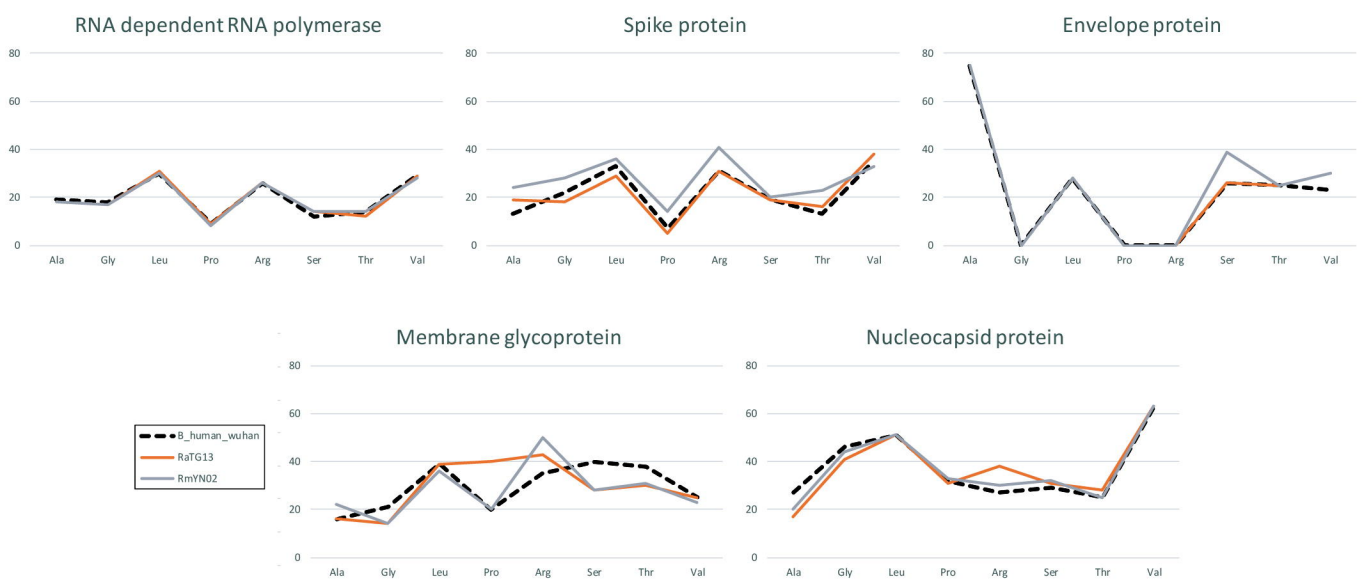
bioRxiv preprint doi: https://doi.org/10.1101/2020.10.12.335521; this version posted October 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

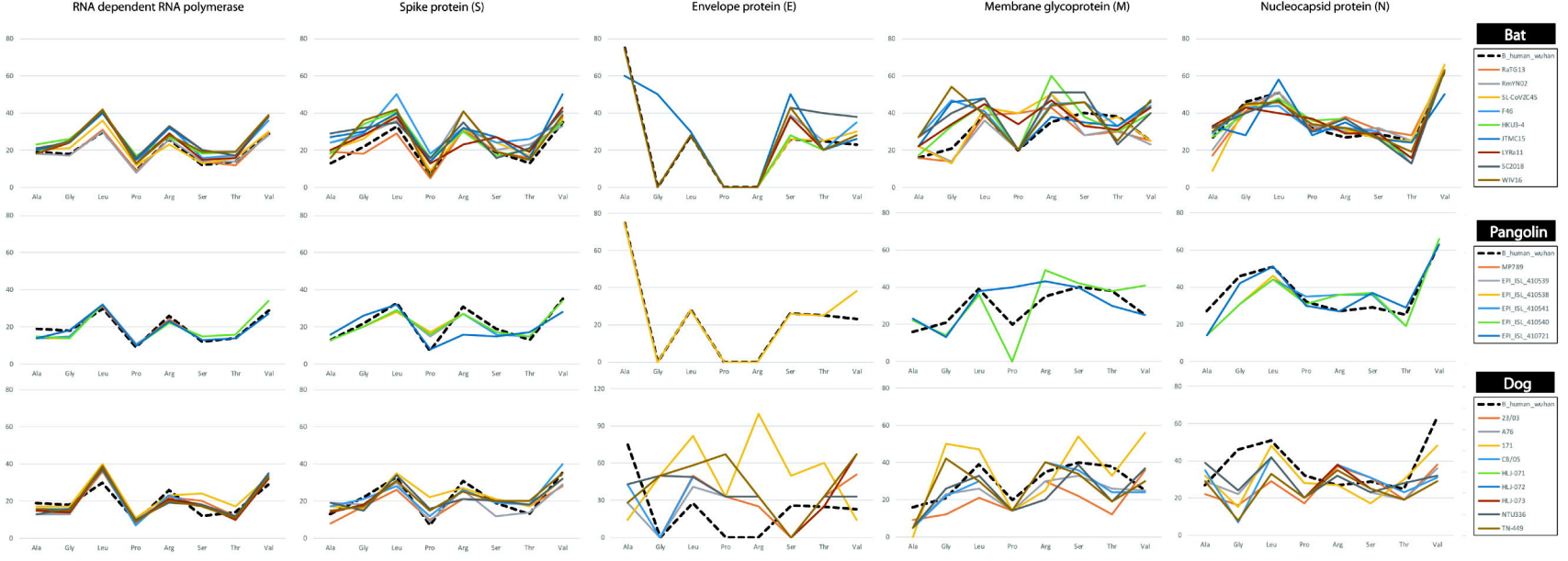
Supplementary figure 3: CUP for E gene for the human SARS-CoV-2 isolates. Eight amino acids with at least four synonymous codons are represented on the X-axis. The percentage of codons ending with G/C for each amino acid is represented on Y-axis.

Supplementary figure 4: CUP for the M gene for the human SARS-CoV-2 isolates. Eight amino acids with at least four synonymous codons are represented on the X-axis. The percentage of codons ending with G/C for each amino acid is represented on Y-axis.

Supplementary figure 5: CUP for the N gene for the human SARS-CoV-2 isolates. Eight amino acids with at least four synonymous codons are represented on the X-axis. The percentage of codons ending with G/C for each amino acid is represented on Y-axis.


80
70
60
50
40
30
20
10
0
Ala
Gly
Leu
Pro
Arg
Ser
Thr
Val
rdrp
S
I
M
N


RNA dependent RNA polymerase
Spike protein
Envelope protein
Membrane glycoprotein
Nucleocapsid protein
80
60
40
20
0
Ala
Gly
Leu
Pro
Arg
Ser
Thr
Val
B_human_wuhan
RaTG13
RmYN02


RNA dependent RNA polymerase
Spike protein (S)
Envelope protein (E)
Membrane glycoprotein (M)
Nucleocapsid protein (N)
Ala
Gly
Leu
Pro
Arg
Ser
Thr
Val
Bat
B_human_wuhan
RaTG13
RmYN02
SL-CoVZC45
F46
HKU3-4
JTMC15
LYRa11
SC2018
WIV16
Pangolin
B_human_wuhan
MP789
EPI_ISL_410539
EPI_ISL_410538
EPI_ISL_410541
EPI_ISL_410540
EPI_ISL_410721
Dog
B_human_wuhan
23/03
A76
171
CB/05
HLJ-071
HLJ-072
HLJ-073
NTU036
TN-449